# RETHINKING THE PRESENTATION OF RESULTS FROM WEB SEARCH

Rahul Singh* Ya-Wen Hsu Wen-Cheng Sun Dil Chitaure Liu Yan
*Department of Computer Science, San Francisco State University, San Francisco, CA 94132*
**rsingh@cs.sfsu.edu {logoin, jasonsun, chitaure, yliu0902}@sfsu.edu*

**ABSTRACT**

*This paper describes a novel approach to the presentation of results from web-search by utilizing the semantic correlations amongst the retrieved results as well as their spatio-temporal characteristics. The fundamental goals of this research are threefold; first it seeks to facilitate easier access to information than what is provided by the current paradigm of listing the retrieved hits in order of relevance. Second, and more importantly, it attempts to recognize, retain, and present context within the retrieved data. Such support is essential for assimilation of the information obtained as the consequence of a query. Finally, it seeks to support both direct search behavior as well as more exploratory search strategies. In this paper we describe the algorithmic underpinnings, based on Latent Semantic Analysis, of our approach as well as the strategy for spatio-temporal information display and querying. Examples involving different types of queries and comparative experiments illustrate the idea and verify its efficacy.*

## 1. INTRODUCTION

The search for electronic information is a complex process whose significance has heightened with the evolution of the internet and the world-wide-web. The unique complexity of the web as an information repository is highlighted by a multitude of factors which include among others, the volume and diversity of information, lack of a rigorous data model, asynchronous changes, and the polysemous nature of many terms used for describing information. These factors combine to create a unique set of challenges for information retrieval on the web.

In this context, web search engines have evolved to become powerful retrieval tools that are used ubiquitously. Designing a successful web-search approach is predicated on handling three major classes of technical issues. The first of these involves the design of techniques capable of indexing all available information as completely as possible. The second set of problems relates to the choice of the information model employed by the search engine and the specific retrieval algorithms used. The final group of issues relates to the query formulation and result presentation capabilities supported by a specific search strategy.

At the current state-of-the-art, the first set of issues has been the focus of significant research and development which has lead to the design of efficient crawlers running on large computer farms. Likewise, the issues of information modeling and retrieval algorithms for the web continue to be actively investigated. We refer the reader to [3] (and references therein) for a review of the developments in this area. Notably, investigators have also reported the use of unified multimedia data models (as opposed to the web-document model) to structure and present information on the web [4]. Attempts have also been made to improve query formulation during web-searches. They have generally focused on improving keyword search by using natural languages [8], metadata [18], or context [7].

Unlike the aforementioned problems, techniques for presentation of results from web searches have tended to evolve slowly. Notwithstanding approaches like page ranking [2], users are still confronted with a large number of hits presented across multiple pages. Current paradigms for presenting results provide little information on how the hits are related or what the result set contains beyond the links displayed on the first page. This forces users to sift through large volumes of data to locate and assimilate the desired information.

We believe that re-thinking the presentation of results from web search can be a critical step in ameliorating these problems. Any attempt towards this needs to recognize and address the following, often interrelated challenges:

- *Supporting data context by recognizing the semantic correlations between retrieved links*. Recognizing and bringing forth the correlations in the data can help orient users in their exploration of the information space. It also aids in information assimilation by underlining the semantic links underlying the data. From a design perspective, information about such correlations can be utilized for reducing presentation clutter by grouping related items.
- *Provision of mechanisms that providing an overview of the entire (retrieved) information content*: Such an ability can aid users in rapidly forming a broad view of the informational diversity. It can also be used for supporting broad-to-focused search strategies/behavior.

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

- *Supporting spatial-temporal characteristics of the information:* Spatial-temporal relationships are inherent to information from the real world. They can form the basis of more complex relationships such as cause-effect, often essential for assimilation of information and discerning patterns in it. Spatio-temporal indexing and presentation have also been shown to significantly improve the complexity of locating data [14, 15]
- *Supporting different user strategies for locating information:* User strategies can involve direct search to look for specific information or exploratory search. Both these strategies have been shown to employ localized or situated navigation [16]. An alternate presentation of search results should therefore facilitate both exact search and search by exploration. In part, such facilitation would depend on resolving many of the aforementioned issues such as support for overview and context along with efficient techniques for data access.

## 2. OVERVIEW OF SOLUTION METHODOLOGY AND DISTICTIONS FROM PRIOR RESEARCH

As can be envisaged from the aforementioned discussion, a key component of the problem complexity lies in determining the semantic correlations between different web pages retrieved as the consequence of a query. A plausible approach to this issue can be based on determining the similarity of content between the retrieved pages. Such an analysis has to account for the media-types through which the content of the web page is expressed. In this regard, textual data is a relatively common media type across a wide spectrum of web pages. Furthermore, textual content is often critically employed in web pages for conveying information. A key thrust of our research therefore is on determining the similarity of web pages based on analyzing the similarity of their textual content. Additionally, we seek to identify spatio-temporal relationships that exist in the extracted information. At the current stage of our research, we parse the content of each web-page to identify (if available), the geographical location information and any associated time information. It should be noted that such spatio-temporal information may either not be available or even applicable for all web-content. However, when such information does exist, the ability to explicitly capture and display it can be of help in understanding complex relationships underlying the data, such as spatial, temporal, and spatio-temporal co-occurrence, spatial and temporal information distribution, and temporal order. Discerning such patterns can help in reasoning about cause-effect and evolutionary relationships. Additionally, recognizing spatial and temporal characteristics of the data allows us to query and retrieve such information using natural spatial and temporal cues. This research (see the experimental section), as well as prior investigations [15] show that such spatio-temporal queries can support extremely intuitive and rapid access to pertinent data.

Our approach consists of two key steps. First, we analyze the textual-content similarity of retrieved web-pages using Latent Semantic Analysis (LSA). LSA is a theory and method for extracting and representing contextual-usage meaning of words by statistical computations and has been extensively used in text analysis [6]. Following the basic approach of LSA, we map the web-documents retrieved as the result of a query to a low-dimensional subspace where they are clustered. Each cluster typically represents a set of correlated documents. Key terms within each cluster are then used to label it. At the end of this step, the traditional linear presentation of the result list is replaced by presenting each link in a group which contains other related links. For each group, the key terms provide a high-level overview of its content. Second, each of the links in a cluster are parsed to extract location and time data. An interactive map and timeline are then used to display this information. In addition to visualization, the spatio-temporal display supports direct interactions with the data. For instance, users can select a geographical location on the map to identify the retrieved hits related to that area.

Owing to the importance of the web, research from both academic and industrial initiatives is starting to address the problem presented in this paper. In [12], position sensitive word-based and TF-IDF-based clustering is proposed for user selected links. The internet search engine Northern Light [10] attempts to organize results by clustering them into categories predefined in library sciences. The Vivisimo search engine [17], provides a generic text-based clustering capability using a heuristic that looks for result subsets having concise readable descriptions. Grokker [5] is another commercial product that provides interactive hierarchical clustering of search results. While many specific technical distinctions can be pointed out between our approach and prior work in this area, the fundamental distinction lies in our attempt not just to cluster the results based on textual similarity, but also to identify the spatio-temporal distribution of the results, and support data context and overview.

## 3. PROPOSED METHOD

***Preprocessing:*** First, commonly occurring words in the retrieved texts are eliminated using a 670-word stop-list. A term frequency table $F$ is then constructed. For each term in the term frequency table, the row-entropy is computed as follows:

$$E(j) = \sum_{i=0}^{n} P(i,j) * \log(\frac{1}{P(i,j)}) \quad (1)$$

Where n is total number of documents to be clustered, and $P(i,j)$ is the occurrence probability of term $j$ in document $i$. Following this, each element $F(i,j)$ in the term frequency table is normalized using the following formula:

$$F(i,j) = \frac{\log(F(i,j))}{E(j)} \quad (2)$$

***Latent Semantic Analysis***: The processed frequency table is decomposed using the singular value decomposition (SVD) [1]. A dimensionality reduction step is then executed and a new approximation of the frequency table is computed. Using this approximation, a correlation matrix for the retrieved web-pages is determined.

***Clustering***: Our clustering algorithm is developed by utilizing the data from the correlation table created during the previous step and a predefined correlation threshold. We use a k-medoids clustering method, which defines the cluster centroid as the document most similar to the rest of the documents in the cluster and separates all documents into k clusters. The algorithm begins with one cluster where all documents are placed. If all correlation values are larger than the threshold, there is no need to further group these documents. Otherwise, the cluster number k is incremented by 1 and all documents are clustered into k groups using k-medoids approach. The above step is iterated until all clusters have documents with correlation values above the threshold and each cluster is not similar with each other.

***Cluster labeling***: We implement a labeling algorithm based on using TF-IDF[9, 13] (Term Frequency- Inverse Document Frequency). The formula we used had been proved by Saltonto and Buckley[13] to have better result in short query vectors. The full TF-IDF document term weight formula [13] is:

$$tf \times \log\frac{N}{n} \Big/ \sqrt{\sum tf \times \log\frac{N}{n}} \quad (3)$$

Here, *tf* stands for term frequency table. An *idf* factor is computed as log N/n, in which N stands for total collection of documents and n is the number of documents which a term is assigned. Finally, a normalization factor is applied, which treats all relevant documents equally regardless of the length of document vectors. The best term we want to extract should have high term frequency and low overall collection frequency.

***Spatial-Temporal information extraction***: The goal in this step is to identify (available) location and/or time characteristics of the data. Our approach is based on using the OpenMap library [11]. A comprehensive directory of location names and latitude-longitude information is used to parse each web page and identify the spatial characteristics of the information contained in it. Similarly, the pages are parsed to extract any available time information that can be used to represent the data on a timeline supporting year-to-month temporal granularity.

***Spatio-Temporal display and query***: The map and the timeline are cursor-location sensitive. This allows users to interact with the data. For example, queries to identify information from a specific geographic region or temporal interval can be issued by selecting it. Figure 1 shows the user-interface depicting document clusters along with spatio-temporal distribution of the information resulting from a query on "earthquakes" (see following section for more details).

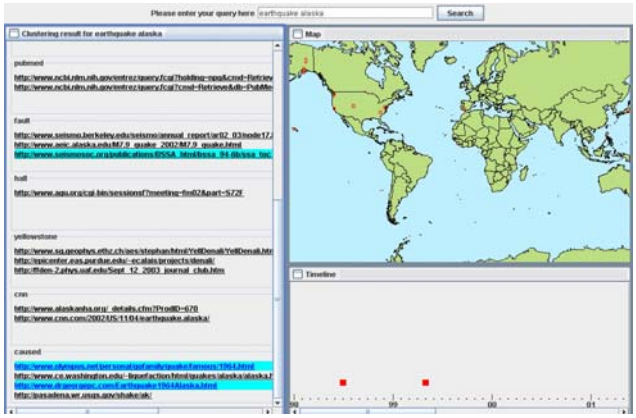


**Figure 1: Spatial display of search results using the "earthquake Alaska" query**

## 4. EXPERIMENTAL RESULTS

We present two set of experiments to demonstrate the applicability of the proposed approach. The first experiment seeks to validate the applicability of pre-processing and latent semantic analysis to group semantically related web pages. We used "ant" as query term and randomly selected 8 of the first 20 documents returned by Google. A synopsis of the content of these documents is shown in Table 1. A manual analysis indicates that the following links are related in terms of their content: {0, 17} (apache projects and tools) and {8, 10, 14, 18, 19}(ant species), while {3} (Ant-zen audio and visual arts) exemplifies a distinct cluster. Here and in the discussion below, numbers inside second brackets indicate the document numbers.

| Document Number | Description |
|---|---|
| 0 | Apache ant project |
| 17 | Apache ant is open-source java-based build tool |
| 8 | Antbase is providing for the first time access to all the ant species of the world. |
| 10 | Ant colony cycle |
| 14 | Carpenter ant: habitat, life cycle, type of damage, control |
| 18 | Pharaoh ant: identification, life cycle and habits, control and prevention |
| 19 | Fire ant research and management project |
| 3 | Ant-zen audio and visual arts, industrial music |

**Table 1: Synopsis of the contents of some of the web-pages**

In Table 2, the correlations between these documents after latent semantic analysis are shown. The average correlation for the cluster {0, 17} is 0.98, and that for {8, 10, 14, 18} is 0.983. As expected, Cluster {19} is on average closer to cluster {8, 10, 14, 18} than to the unrelated clusters {0, 17} or {3}. These results correspond to the manual grouping of the pages, thus demonstrating the applicability of latent semantic analysis to group web pages that are related in terms of their textual content.

The second experiment demonstrates the applicability of spatio-temporal querying to rapidly access information. In it, the goal is to find information on earthquakes in specific parts of the world and at specific times; query-1: 1964 in Alaska, query-2: 1906 in San Francisco, and query-3: 2004 in Indonesia. In query-4, seeks to find the latest score of a game between two baseball teams: Giants and Rockies.

| | 0 | 17 | 8 | 10 | 14 | 18 | 19 | 3 |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | | | | | | | |
| 17 | 0.98 | 1 | | | | | | |
| 8 | 0.67 | 0.55 | 1 | | | | | |
| 10 | 0.66 | 0.55 | 0.99 | 1 | | | | |
| 14 | 0.64 | 0.52 | 0.99 | 0.99 | 1 | | | |
| 18 | 0.77 | 0.66 | 0.98 | 0.98 | 0.97 | 1 | | |
| 19 | 0.80 | 0.69 | 0.98 | 0.97 | 0.97 | 0.99 | 1 | |
| 3 | 0.34 | 0.25 | 0.14 | 0.13 | 0.13 | 0.24 | 0.24 | 1 |

**Table 2: Correlations between web-pages after LSA**

This example typifies information search characteristic to a plethora of significant real-world applications from disaster management to travel. Using the proposed approach, the geographical and time distribution of the information can be immediately discerned (Figure 1, documents in selected region or time are highlighted). Thus, by selecting the regions and time of interest, appropriate information can be directly accessed. In Table 3, we present the number of mouse clicks required to access the information using the proposed approach and the presentation strategies supported by Google, Yahoo, and Vivisimo using the following modified queries; query-1: "Earthquake Alaska", query-2: "Earthquake San Francisco", query-3: "Earthquake Indonesia" and query-4: "Baseball San Francisco 2005". We note that the proposed approach leads to the fastest access (lowest number of clicks) even in cases where queries in other system were modified to explicitly include the region or time of interest.

| Approach | Query-1 | Query-2 | Query-3 | Query-4 |
|---|---|---|---|---|
| Yahoo | 17 | 13 | 7 | 5 |
| Vivisimo | 7 | 30 | 7 | 11 |
| Google | 10 | 13 | 6 | 5 |
| Proposed Approach | 5 | 6 | 4 | 2 |

**Table 3: Complexity (number of clicks) for accessing specific data.**

## 5. CONCLUSIONS

In this paper, we present our research on exploring alternate paradigms for displaying results from web-search engines. The primary motivations behind this research are: (1) Meeting the goal of complex information assimilation and (2) Supporting efficient access to information as well as supporting exploratory information search. Our approach to this problem has focused on determining the similarity (using latent semantic analysis) between the textual content of retrieved pages. This information is then used to cluster topically related documents. Additionally, we parse the retrieved pages to extract available spatio-temporal information. This is then presented through an interactive interface both for purposes of visualization as well as interactions with the retrieved data. The proposed approach makes a cardinal departure from the current state-of-the-art through its support for context, information overview, and support of spatio-temporal information display and interaction. Given the increasing complexity of information being presented through the web, we believe, approaches such as ours, can be of substantial use in information query, display, and assimilation.

## 6. REFERENCES


[1] M. W. Berry, S. Dumais, G. W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval", SIAM Review 37:4 1995

[2] Brin, S., And Page, L. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In Proceedings of the 7th International World Wide Web Conference, 1998.

[3] E. Dyson, "The Search for Structure", Esther Dyson Monthly Report, Vol. 21, No. 1, January 2003

[4] B. Gong, R. Singh, and R. Jain, "ResearchExplorer: Gaining Insights Through Exploration in Multimedia Scientific Data", Proc. 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 7 – 14, 2004

[5] Grokker:www.grokker.com

[6] T. Lander, P. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis", Discourse Processes, Vol. 25, pp. 259-284, 1998

[7] S. Lawrence, "Context in Web-Search", IEEE Data Engineering Bulletin, Vol. 23, No. 3, pp. 25 – 32, 2000

[8] J. Lin et al., "The Role of Context in Question Answering Systems", Proc. CHI 2003

[9] Y. Liu, V. Dasigi, B.J. Cliliax, K.Borges , A.Ram, Navathe, B. Sharmant and R. Dingledine, "Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering" Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference 2004

[10] G. Notess, "Review of Northern Light", 2002, http://www.searchengineshowdown.com/features/nlight/review.html

[11]] openMap:http://openmap.bbn.com

[12] D. Radev and W. Fan, "Automatic Summarization of Search Engine Hit Lists", ACL Workshop on Recent Advancements in NLP and IR, 2000

[13] G. Salton, and C. Buckley,"Term Weighting Approaches in Automatic Text Retrieval" Technical Report: TR87-881 1987

[14] S. Shekhar, S. Chawla, "Spatial Databases: A Tour", Prentice Hall, 2003

[15] R. Singh, R. Knickmeyer, P. Gupta, and R. Jain, "Designing Experiential Environments for Management of Personal Multimedia", ACM Multimedia, 2004

[16] J. Teevan, C. Alvarado, M. Ackerman, and D. Krager, "The Perfect Search Engine is Not Enough: A study of Orienteering Behavior in Directed Search", Proc. CHI 2004

[17] Vivisimo:www.vivisimo.com

[18] K-P. Yee, K. Swearingen, K. Li, and M. Hearst, "Faceted Metadata for Image Searching and Browsing", Proc. CHI, 2003